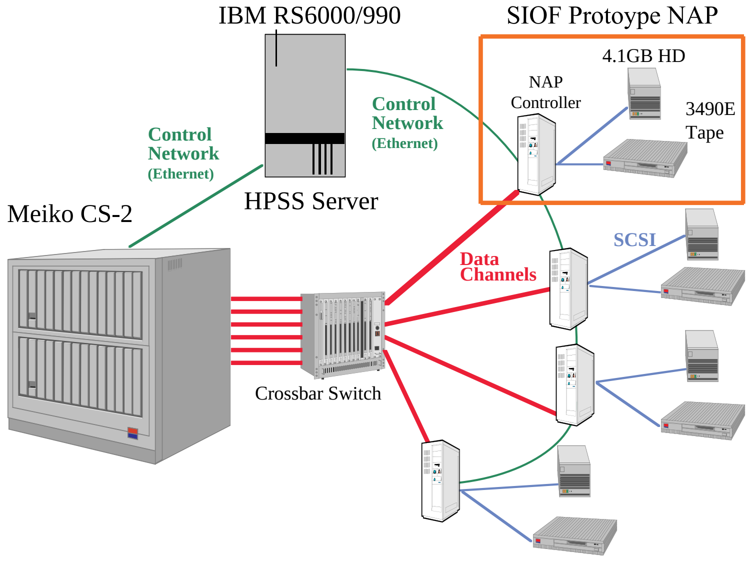
The Scalable I/O Facility

Breaking the I/O Bottleneck

University of California
Lawrence Livermore
National Laboratory



Scalable I/O Facility architecture.

Objective

The Scalable I/O Facility (SIOF) project at Lawrence Livermore National Laboratory (LLNL) enables input/output (I/O) performance to scale with the computing performance of modern parallel computing systems and thus achieve true terascale computing. SIOF is an implementation of the High Performance Storage System (HPSS) architecture. The SIOF team is a part of the HPSS consortium and collaborates with other academic and industry groups.

Impact

Scalable I/O performance can eliminate the I/O bottleneck currently being experienced by massively parallel processors (MPPs) and high-performance workstation clusters (SMPs). The SIOF project is developing three technologies to satisfy requirements for scalable and practical input/output performance:

- Portable and parallel Application Programming Interface (API) - MPI-IO.
- Cost effective Network Attached Peripherals (NAPs).
- Network technology in dependence.

The challenges of science and industry driving computing and communications have created corresponding challenges in data and information storage and retrieval. Large commercial and scientific applications are straining storage facilities, a condition compounded by MPPs and high-performance clusters. Most current large-scale storage architectures pass their data through CPUs in a centralized shared storage system. Such systems are reaching economic and technological limitations and can no longer meet performance, capacity, and cost requirements. We need an I/O architecture that can scale with the I/O needs of a network of heterogeneous computational resources, much like the computational power of an MPP scales with the number of processors.

Leveraging LLNL and Collaboration Experience

The SIOF project leverages the work of the National Storage Laboratory (NSL) collaboration, which developed a storage system management software product now available as NSL-Unitree, and the work of the High Performance Storage System Consortium, which features a scalable, distributable architecture using parallel I/O and striping across network-attached devices, improved storage system management, and support for data management applications. The SIOF project also leverages the work of LLNL's Advanced Telecommunications Program (ATP) in applying Fibre Channel (FC) standard technology.

SIOF Enhancements to HPSS

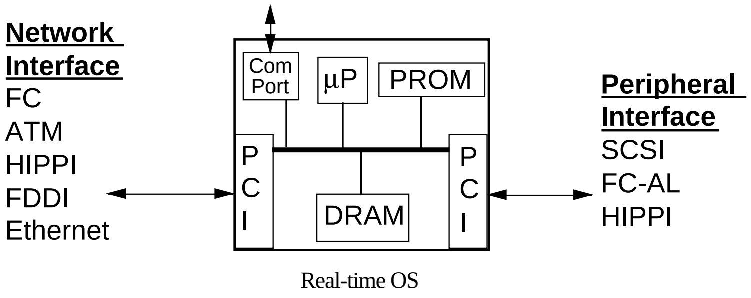
The SIOF project is enhancing the HPSS architecture to provide:

- A simple and portable application interface based on the MPI-IO API.
- A flexible and cost-effective NAP.
- Additional high speed LAN technologies for HPSS.

The SIOF MPI-IO and NAP implementations enhance the capability of HPSS to deliver performance through parallel networking and storage components. The HPSS and SIOF work is driving a new distributed computing paradigm for network-centric storage hierarchies that will allow MPPs and SMPs to achieve teraflop computing.

The HPSS/SIOF Architecture

The goal of SIOF is to demonstrate a network-centered, scalable storage system that supports parallel I/O across the computing environment. This facility employs a crosspoint-switched FC fabric to connect the computing nodes of a Meiko CS-2 parallel processor to disk arrays, parallel tapes, and frame buffers (see diagram). In 1997 we will migrate the facility to our new IBM SP2 parallel processor. The built-in parallelism of this architecture and the scalability of a switched fabric give the parallel applications scalable, direct access to



The next step in Network Attached Periferals. The SIOF NAP Controller—a single board NAP.

peripherals attached to the network. A significant benefit of the architecture is that the I/O bandwidth increases linearly with the number of nodes and peripherals connected to the fabric.

In this SIOF implementation, applications store and retrieve data through the storage system by means of the MPI-IO API. Control messages are sent over an Ethernet network to meet security requirements. Data flows between the storage system and the application via the prototype NAPs over the high-performance network. It is the goal of the SIOF project to transfer the SIOF technologies into commercial products.

The SIOF Project Components

NAP

NAPs make storage resources directly available to computer systems on a network without requiring a high-powered processing capability connected to the storage devices for file services and data transfer. With the NAP approach, a single network-attached control system (e.g., HPSS) can manage access to the storage devices without being required to handle the transferred data. This approach allows the control system to scale to much higher levels while still supplying centralized control for such functions as naming and access control. SIOF work currently under way on our prototype NAPs will increase the per-port channel transfer bandwidth to the maximum available from the CPU, network, and peripherals; decrease the cost of network-attached peripherals to roughly the commodity cost of the peripherals themselves; and provide enhanced security for NAPs (e.g., eliminate the need for a separate control network).

MPI-IO

MPI is a recently emerging standard for message passing in parallel computing systems. Part of the MPI standard is the MPI/IO input/output interface. The SIOF project has implemented the MPI/IO interface for the HPSS system on a Meiko CS-2. Work is under way on porting this implementation to our recently acquired IBM SP2. SIOF staff members are participating in the MPI-IO standards working group and are playing a key role in the efforts to develop MPI-IO into a standard. We will continue to expand and enhance the capabilities of this interface while tracking the developments of the emerging MPI-IO standard. We are also working with other groups implementing the MPI-IO API on other platforms to ensure that all such implementations interoperate and meet the needs for scalable, parallel I/O performance.

Network Independence

The only currently available commercial NAP products are expensive and network media-dependent. They tie specific command protocols to specific network fabrics—e.g., IPI-3 to HIPPI and SCSI-3 to Fibre Channel. The work in the Scalable I/O Facility project is directed toward fabric-independent NAPs. Our current prototype NAP is based on workstation technology and uses TCP/IP for network fabric transport to achieve network independence.

We expect SIOF's NAP Controller prototype to also use TCP/IP to achieve network fabric independence. This network independence is important both to support wide-area network data access where TCP/IP is the only protocol available and also to support the variety of evolving high-performance LAN networking technologies—e.g., ATM, Fast Ethernet, FDDI, Fibre Channel, HIPPI.

Industry Collaboration

SIOF is collaborating with a number of academic and industry groups, as a part of the HPSS Consortium, and by working with the SCSI and Fibre Channel Standard committees in X3T11, the Scalable I/O Initiative, and the National Storage Industry Consortium's Network Attached Storage Device (NASD) working group. The SIOF is being funded by the Advanced Strategic Computing Initiative (ASCI) and Department of Energy (DOE) Defense Programs.

We invite others with interests in seeing input/output systems match the parallel performance potential from MPPs and SMP clusters to contact us or participate in this important high-performance parallel I/O work.

For more information, contact
Kim Minuzzo, SIOF Project Leader,
510-422-2141, <minuzzo1@llnl.gov>

Work performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48